Intrinsic Dimension Estimation

One particular problem in the area of dimension reduction is estimating the intrinsic dimension of the underlying manifold. Indeed, the minimum number of required parameters for describing the observed data is the intrinsic dimension. Moreover, knowing the dimension of the submanifold on which the data lives can be helpful when considering algorithms which search nearest neighbours, since their complexity increases exponentially with the dimension. Given the relevance of the problem, many algorithms have been proposed since the first by Bennett, in 1969. We may group them into three classes: local, global, and pointwise. The local algorithms make use of information contained in sample neighbourhoods, while the global algorithms use the whole data set. The pointwise algorithms can provide either global intrinsic dimension estimates if applied to the whole data set, or local ones when applied to sample subsets.

The ideal intrinsic dimension estimator would satisfy the following conditions: 1) be computationally feasible, 2) be robust to multiscaling, 3) be robust to high dimensionality, 4) have a work envelope (operative range), and 5) be accurate. The first point is for practicality reasons. The second point is related to the dependence of the intrinsic dimension on the scale used to view the data. For example, a three-dimensional data set on a 3-sphere with four-dimensional Gaussian noise would appear three-dimensional at coarse scales, and four-dimensional at fine scales. The third point is relevant since much real-life data in, for instance, the medical field is very high-dimensional. The fourth point means that we should be able to give reliable estimates within some determined range of dimensions of the data. The fifth point is once again for practicality.

We would like to consider the examples of two proposed estimators described by Kleindessner and von Luxburg making use of neighbourhood information but not metric information. The k-Nearest-Neighbours (k-NN) estimator falls under the category of global estimators, but it unfortunately fails the multiscaling, high dimenionality, and work envelope points for the ideal estimator above. The non-metric estimators E_{DP} and E_{CAP} are also global, and by purposefully making use of less information they allow for more possible applications, where the data available does not give distance information. However, the naive estimator E_{DP} is poorly-behaved, while the E_{CAP} estimator seems to have similar or slightly worse performance than commonly used estimators for intrinsic dimension.

Let $\mathscr{Y}_n = \{Y_1, \ldots, Y_n\}$ be n independent and identically distributed random vectors with values in a compact subset of $\mathbb{R}^D$. The $k-$nearest neighbours of Y_i are defined as the closest points to Y_i in $\mathscr{Y}_n$ with respect to Euclidean distance:

$$\arg \min_{\{Y_{j_1},\ldots,Y_{j_k}\}\in\mathscr{Y}_n} \sum d_{\mathbb{R}^D}(Y, Y_i)$$

The k-NN graph G is formed by putting edges between each point in $\mathscr{Y}_n$ and its $k-$nearest neighbours. We begin with an estimator based on the doubling property of the volume of a $d-$dimensional ball:

Take the case of $\mathscr{Y}_n$, and build out of it the k-NN graph G with $V = \{1, \ldots, n\}$ the set of vertices. We direct the graph by choosing a directed edge from i to j if and only if Y_j is among the $k-$nearest neighbours of Y_i with respect to the Euclidean distance. Let $B_{SP}(i, r)$ denote the closed ball with centre i in V and radius $r > 0$ in the graph G with respect to the directed shortest path distance d_{SP}. Denote by the usual $B(y, r)$ the closed ball in $\mathbb{R}^D$ with centre y and radius $r > 0$. Denote also by λ_D the $D-$dimensional Lebesgue measure and by $\eta_D = \lambda_D(B(0, 1))$ the volume of the $D-$dimensional unit ball.

We make use of the doubling property of the volume: if x in $\mathbb{R}^d$ and $r > 0$ are given, then

$$\lambda_d(B(x, 2r)) = 2^d \lambda_d(B(x, r))$$

Therefore, the dimension can be found by

$$d = -\log_2 \frac{\lambda_d(B(x, r)}{\lambda_d(B(x, 2r))}$$

Now consider a sample point Y_i and the balls $B_{SP}(i, 1)$ and $B_{SP}(i, 2)$. When n is large and k is relatively small, the points Y_j in $B_{SP}(i, 2)$ are close enough to Y_i that we may approximate this neighbourhood of Y_i in M as a Euclidean neighbourhood in $\mathbb{R}^D$. Indeed, for a small enough choice of $r > 0$, we will have a correspondence of $B_{SP}(i, 1)$ and $B_{SP}(i, 2)$ with $B(Y, r)$ and $B(Y, 2r)$, respectively. With the neighbourhoods sufficiently small, the density is approximately constant, and defining $L_{DP}(i)$ as follows we have:

$$L_{DP}(i) = \frac{|B_{SP}(i, 1)|}{|B_{SP}(i, 2)|} \approx \frac{nf(Y_i)\lambda_d(B(Y, r))}{nf(Y_i)\lambda_d(B(Y, 2r))} = \frac{1}{2^d}$$

Before taking the logarithm, we first average over some subset of vertices $A \subseteq V$:

$$L_{DP}(A) = \frac{1}{|A|} \sum_{i \in A} L_{DP}(i)$$

and we at last obtain the dimension estimator:

$$E_{DP}(A) = -\log_2 L_{DP}(A)$$

This estimator is consistent, but despite converging to the intrinsic dimension d in probability, it requires a very large sample size to give accurate results. Moreover, it often underestimates the intrinsic dimension. We therefore turn towards a second estimator which does not rely on exact distance calculations, but using a different geometric principle.

Fix x, y in $\mathbb{R}^d$ with $d_{\mathbb{R}^d}(x, y) = r$. The set $B(x, r) \cap B(y, r)$ is the union of two spherical caps with height $r/2$ of a ball of radius r. The volume of such a cap is known to be

$$\frac{1}{2}\eta_d r^d I_{3/4}\left(\frac{d+1}{2}, \frac{1}{2}\right)$$

where $I_z(a,b)$ is the regularized incomplete beta function. By taking a quotient of volumes again, we arrive at

$$\frac{\lambda_d(B(x,r)\cap B(y,r))}{\lambda_d(B(x,r))} = I_{3/4}\left(\frac{d+1}{2},\frac{1}{2}\right)$$

and we denote this quantity by $S(d)$. We could recover d from $S(d)$ by inverting the function. Now, to implement it with a sample set $\mathscr{Y}_n$, let us fix a point Y_i and replace $B(Y_i,r)$ by $B(i,1)$. Notice that $|B(i,1)| = k+1$ in the denominator. For the numerator, we want to find a vertex j_0 so that Y_{j_0} is on the boundary of $B(Y_i,r)$. Note also that when we compute $|B_{SP}(i,1)\cap B_{SP}(j,1)|$, it becomes smaller as we increase the distance from Y_i to Y_j, and so we can pick a minimizing vertex j_0 from the vertices j connected to i, yielding:

$$L_{CAP}(i) = \min_{j\in V: i\to j} \frac{|B_{SP}(i,1)\cap B_{SP}(j,1)|}{k+1} \approx S(d)$$

and we may again estimate d by taking the inverse. As above, we average first over some $A\subseteq V$ to get $L_{CAP}(A) = \frac{1}{|A|}\sum_{i\in A} L_{CAP}(i)$, and

$$E_{CAP}(A) = S^{-1}(L_{CAP}(A))$$

Both estimators $E_{DP}(A)$ and $E_{CAP}(A)$ are statistically consistent for any random choice of A. Moreover, their variance decreases approximately as $1/|A|$ if A is chosen uniformly, randomly, and without replacement. However, the new $E_{CAP}(A)$ estimator converges much faster in practice.

Let us make a few comments about the discrepancies in how well these two estimators work. First of all, in both cases, we are inverting some function to obtain the intrinsic dimension estimate. However, notice that whenever $f(x)$ is approximately flat, small deviations in $f(x)$ will lead to large deviations in x. And indeed, the function $1/2^d$ is approximately flat more often than the function $S(d)$. Second of all, in the case of $E_{DP}(A)$, we are using the correspondence of $B_{SP}(i,1)$ and $B_{SP}(i,2)$ with $B(Y,r)$ and $B(Y,2r)$. For the second pairing in particular, $B_{SP}(i,2)$ is the union of $B_{SP}(i,1)$ and the balls $B_{SP}(j,1)$ with $i\to j$. That is to say, the ball $B_{SP}(i,2)$ corresponds to the union of $B(Y_i,r)$ and balls $B(Y_j,r)$. Although this works out in the limit, if we choose finite values of k and n this is not a good approximation. We are filling the larger ball only partially, which leads to the underestimate. This is even worse in higher-dimensional cases, where almost all the volume is concentrated near the boundary.

Bibliography

[1] Block, A., Jia, Z., Polyanskiy, Y., Rakhlin, A., *Intrinsic Dimension Estimation Using Wasserstein Distance.* Journal of Machine Learning Research, Vol 23, 1-37 (2022). http://jmlr.org/papers/v23/21-1483.html

[2] Camastra, F., *Data dimensionality estimation methods: a survey.* Pattern Recognition, Vol 36, Issue 12, 2945-2954 (2003). ISSN 0031-3203

[3] Costa, J. A., Girotra, A., Hero, A. O., *Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs.* IEEE/SP 13th Workshop on Statistical Signal Processing, Bordeaux, France, 417-422 (2005). doi: 10.1109/SSP.2005.1628631

[4] Kleindessner, M., Luxburg, U., *Dimensionality estimation without distances.* Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research, Vol 38, 471-479 (2015)

[5] Pettis, K. W., Bailey, T. A., Jain, A. K., Dubes, R. C., *An Intrinsic Dimensionality Estimator from Near-Neighbor Information.* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAMI-1, No. 1, 25-37 (Jan. 1979), doi:10.1109/TPAMI.1979.4766873

[6] Qiu, H., Yang, Y., Li, B., *Intrinsic dimension estimation based on local adjacency information.* Information Sciences, Vol 558, 21-33 (2021). ISSN 0020-0255,https://doi.org/10.1016/j.ins.2021.01.017